Metagenome Assembly & Binning

also sometimes called *de novo* metagenome analysis

STAMPS 2022

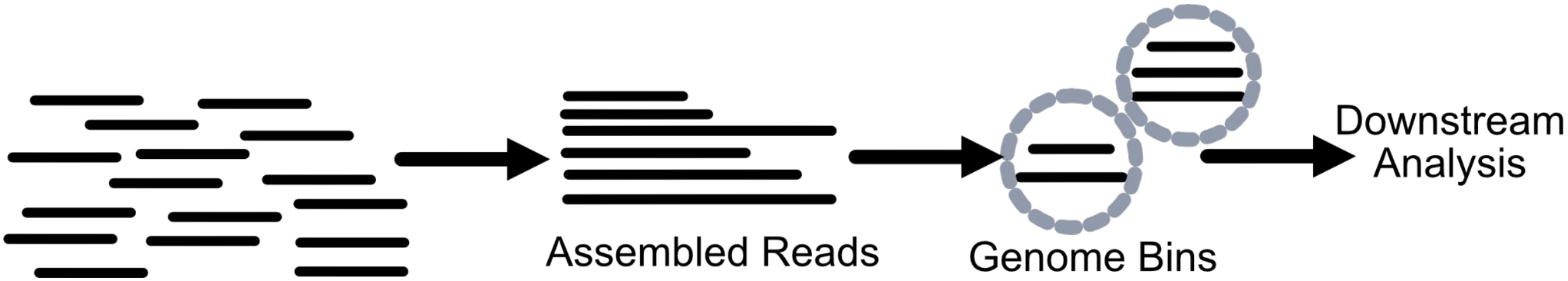
Taylor Reiter, PhD

Metagenome Assembly & Binning mostly for short reads

also sometimes called *de novo* metagenome analysis

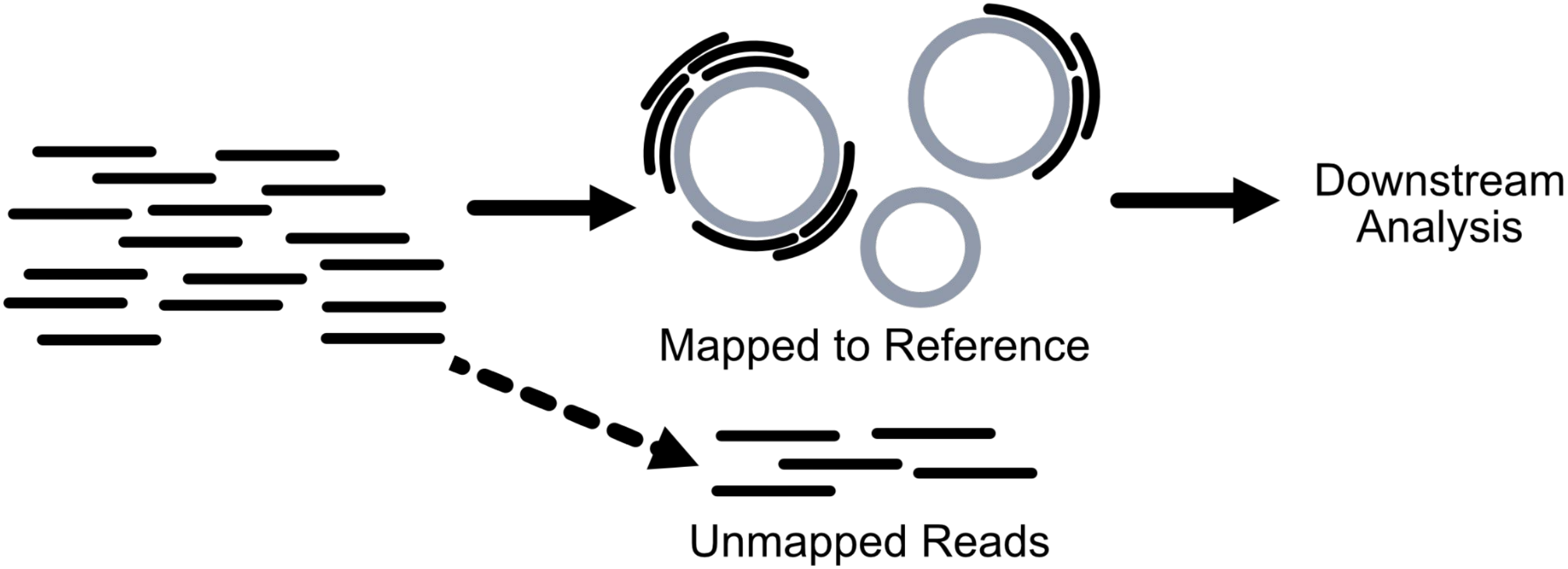
STAMPS 2022

Assembly & Binning

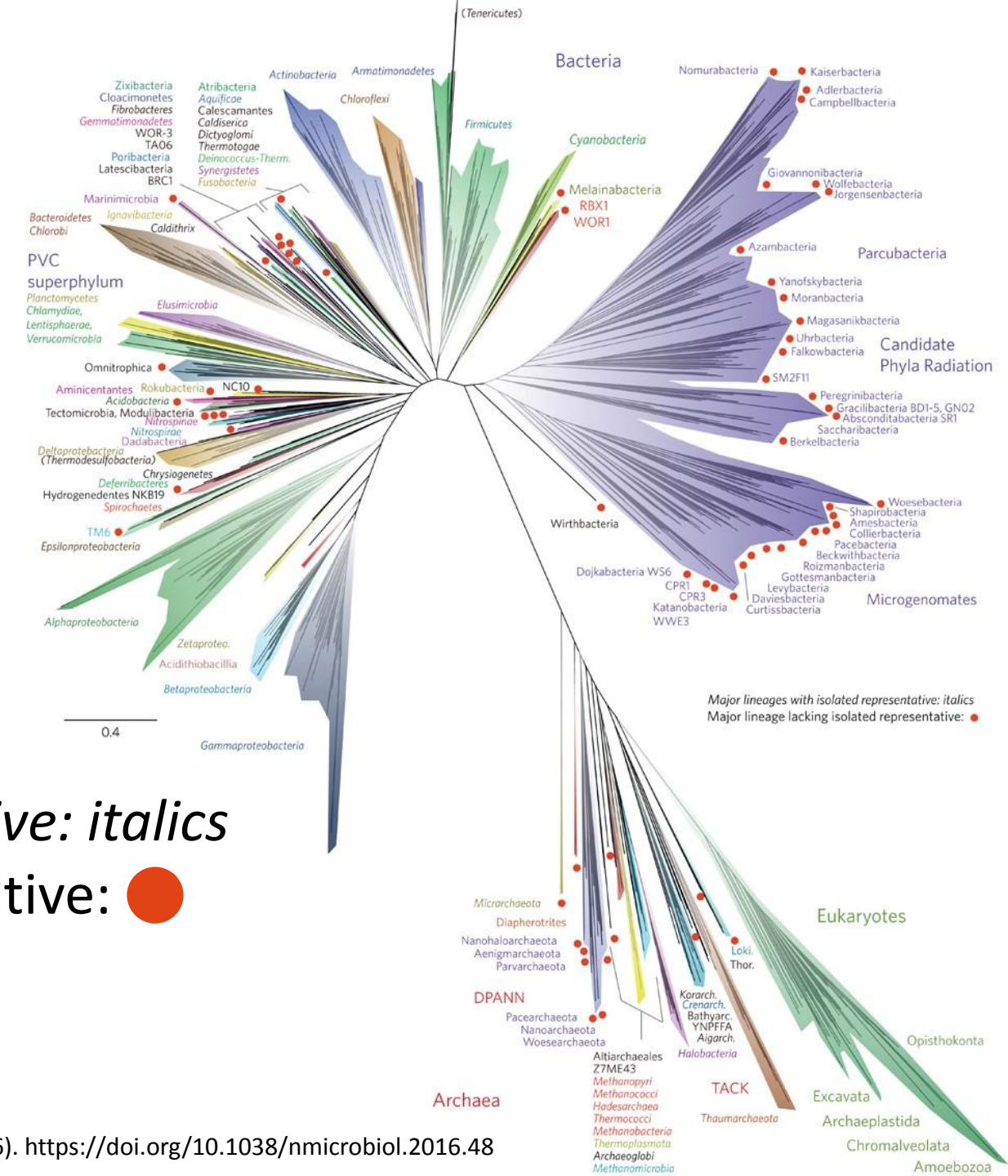


Why do we do *de novo* metagenome analysis?

Why do we do *de novo* metagenome analysis: reference databases are incomplete and we have to do something



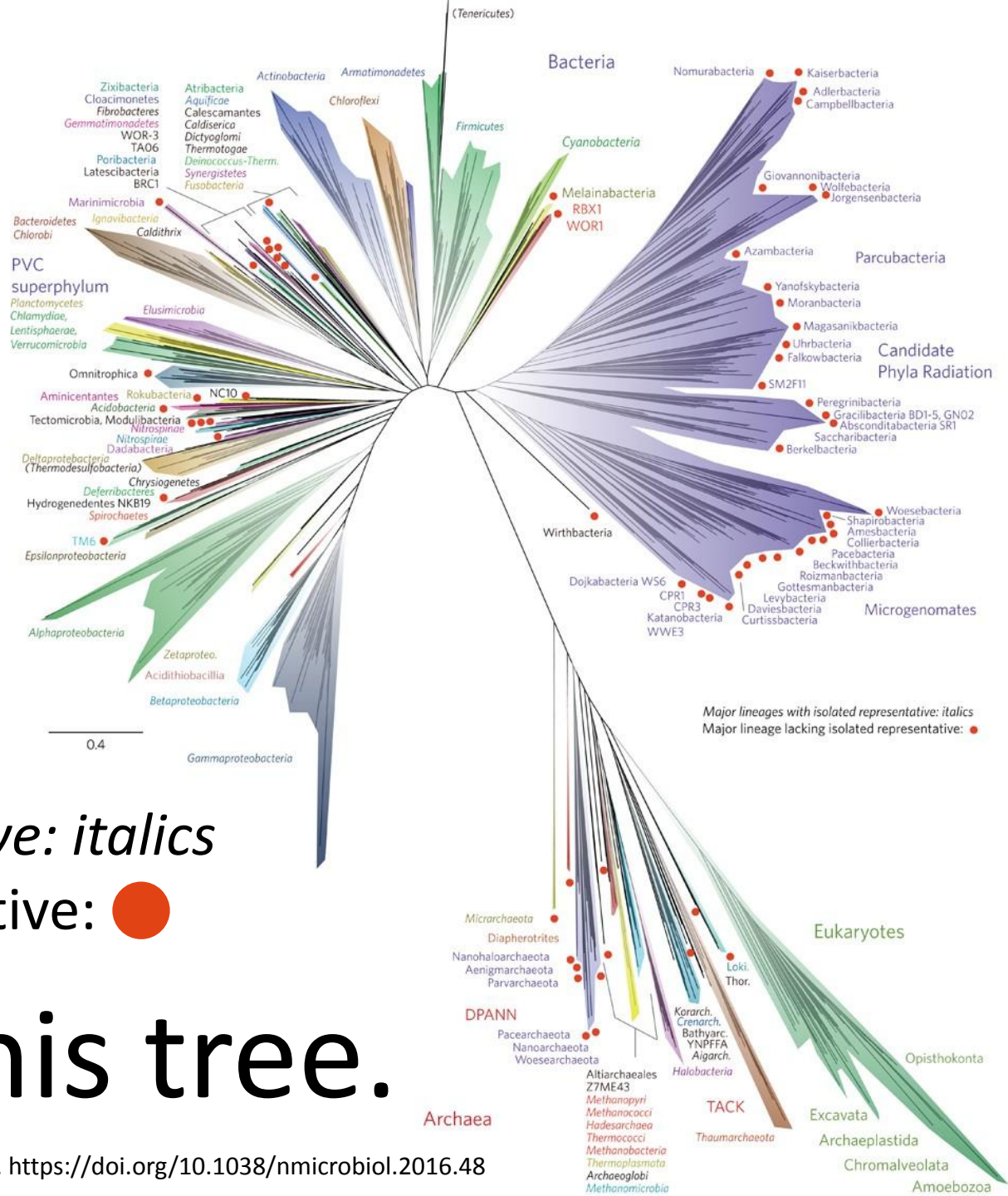
What has *de novo* metagenome analysis given us?



Major lineages with isolated representative: italics

Major lineage lacking isolated representative: ●

Hug, L., Baker, B., Anantharaman, K. *et al.* A new view of the tree of life. *Nat Microbiol* **1,** 16048 (2016). https://doi.org/10.1038/nmicrobiol.2016.48



Major lineages with isolated representative: italics

Major lineage lacking isolated representative: ●

There are **68** ● on this tree.

Hug, L., Baker, B., Anantharaman, K. *et al.* A new view of the tree of life. *Nat Microbiol* **1,** 16048 (2016). https://doi.org/10.1038/nmicrobiol.2016.48

How does assembly work?

It was the best of times, it was the worst of times

Common assembly strategies: **greedy method**

it_was_the_

was_the_wor

orst_of_time

mes_it_was_

Common assembly strategies: **overlap layout consensus method**

it_was_the_

was_the_best

he_best_of

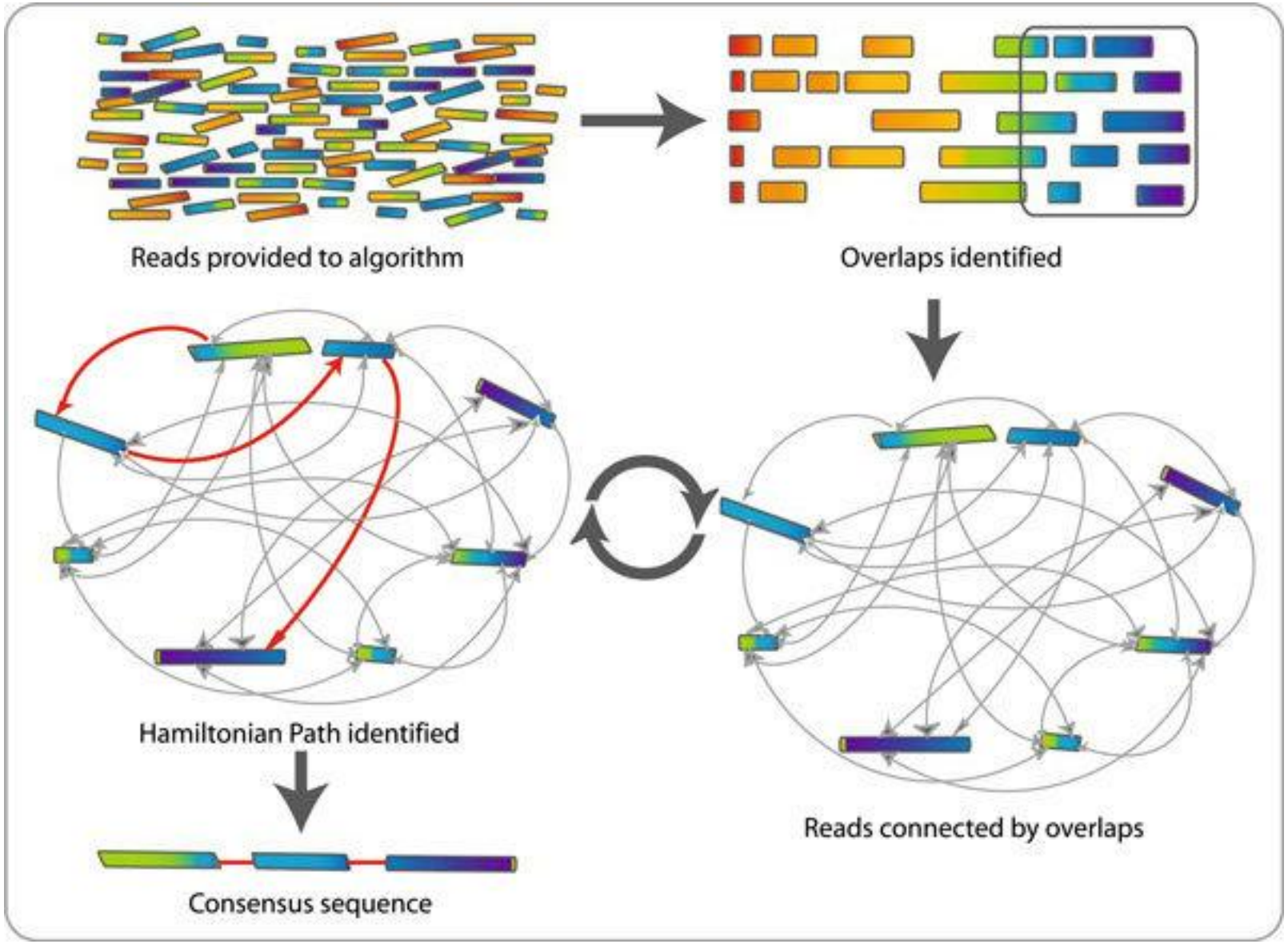
st_of_time

mes_it_was_

Common assembly strategies: **overlap layout consensus method**

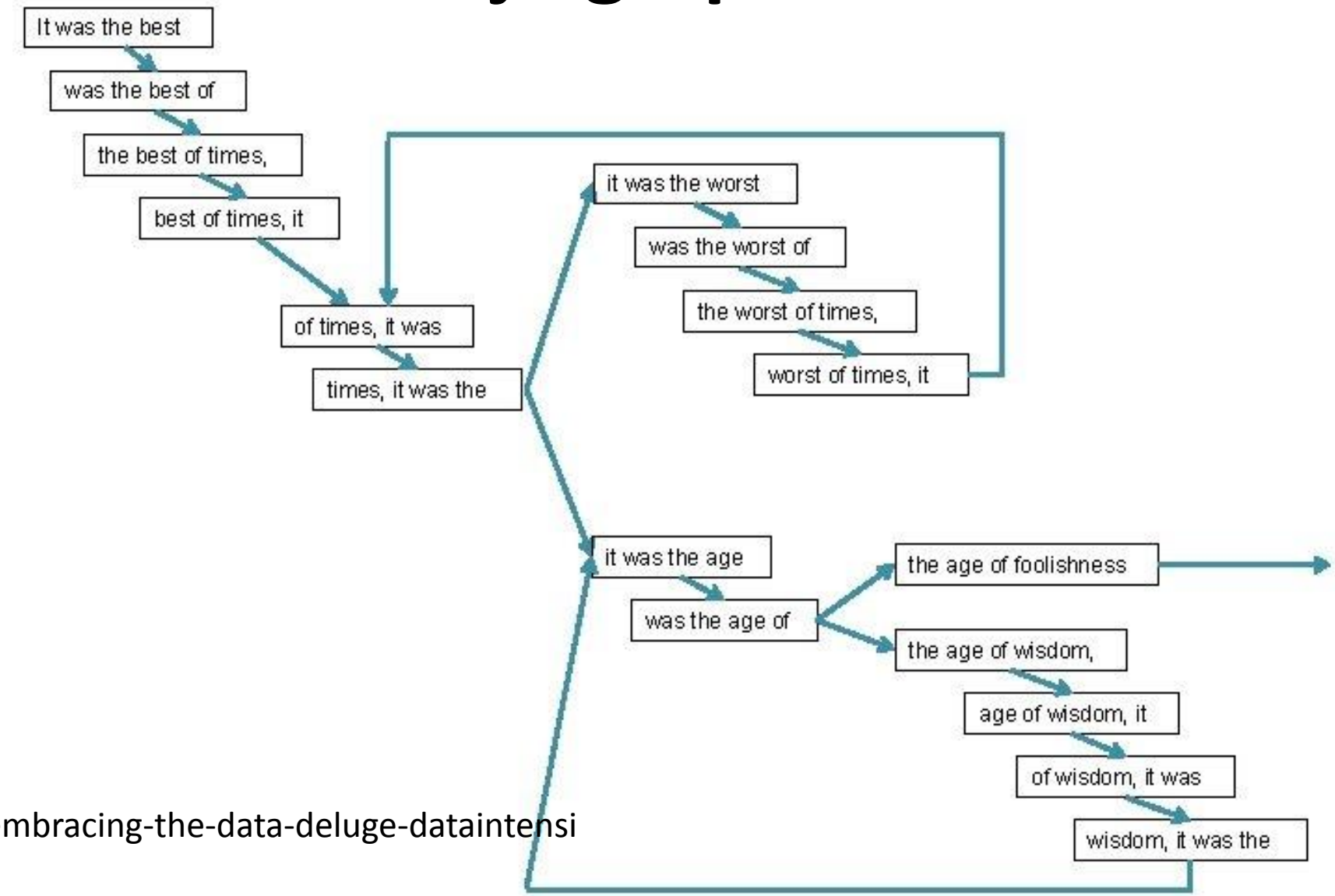
it_was_the_
was_the_best
he_best_of
st_of_time
mes_it_was_

it_was_the_
was_the_wors
he_worst_o
st_of_time



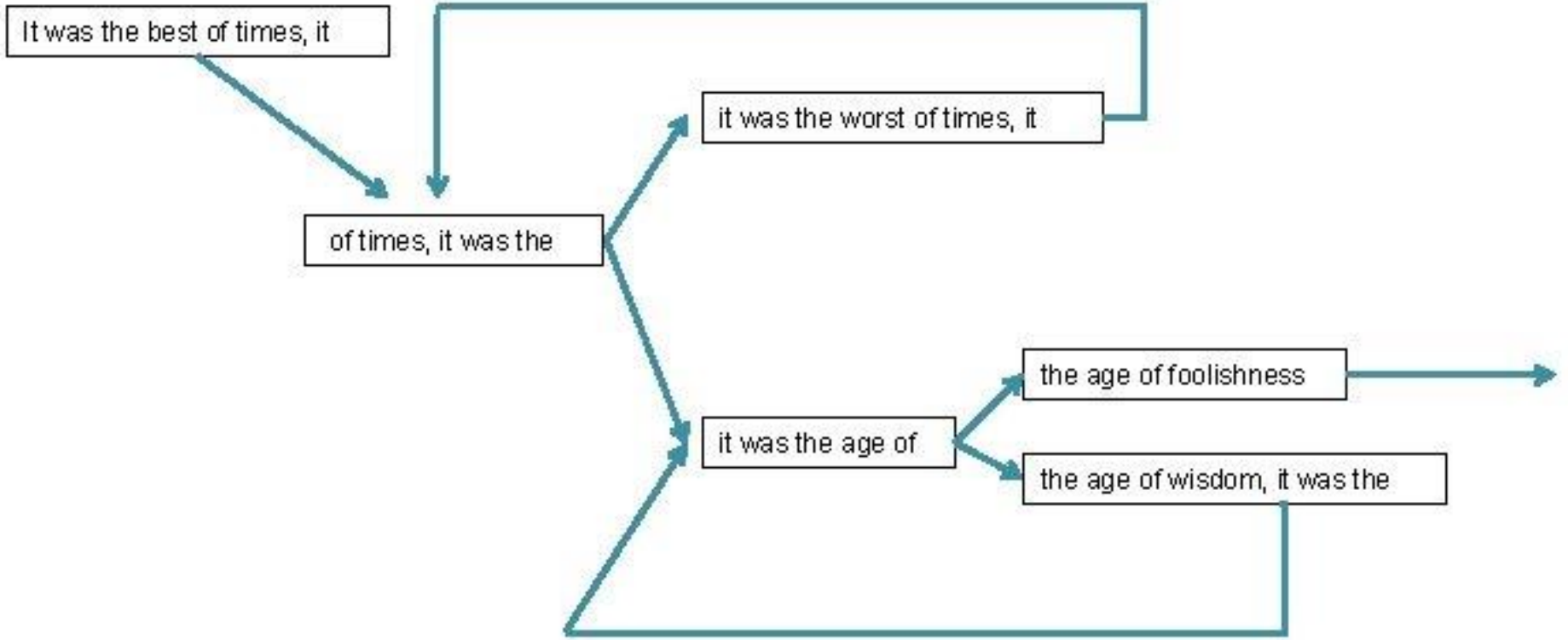
Commins, J., Toft, C. & Fares, M.A. Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects. *Biol Proced Online* **11,** 52 (2009). https://doi.org/10.1007/s12575-009-9004-1

Common assembly strategies: **de Bruijn graph methods**



https://slidetodoc.com/embracing-the-data-deluge-dataintensive-computing-for-the/

Common assembly strategies: **de Bruijn graph methods**



https://slidetodoc.com/embracing-the-data-deluge-dataintensive-computing-for-the/

Tools that do assembly not an exhaustive list

Overlap layout consensus

- ?
- Long read assemblers?

Greedy

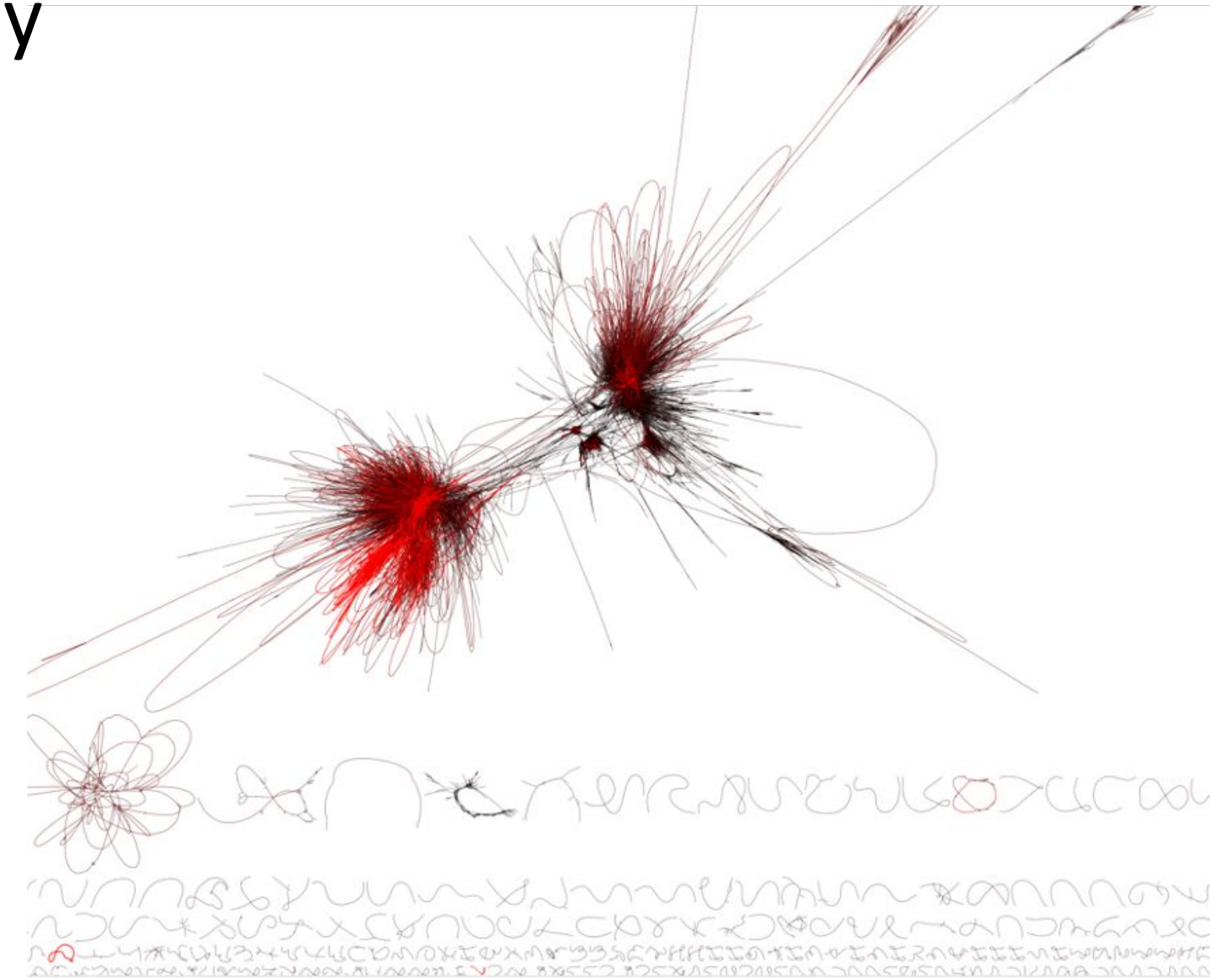
- PLASS

de Bruijn Graph

- (meta)SPAdes
- Megahit
- IDBA-UD
- MetaVelvet
- Ray Meta

Metagenome assembly graphs in the wild

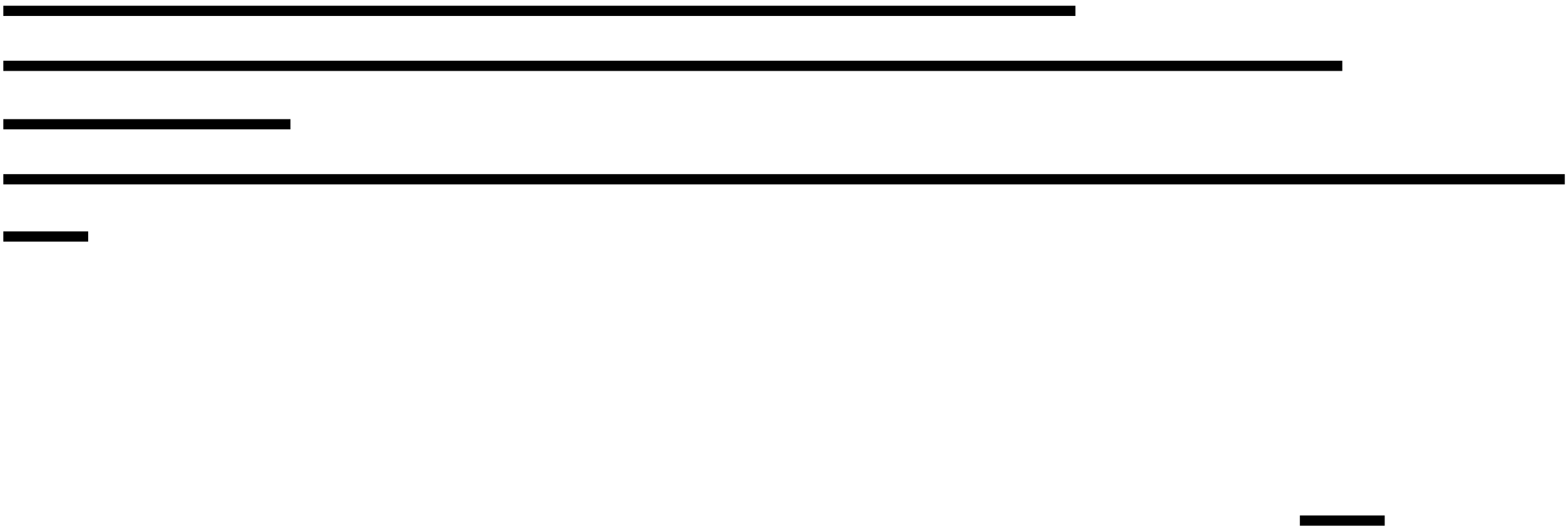
Mouse gut metagenome



@SilasKieser https://twitter.com/SilasKieser/status/1308752555795779585/photo/1

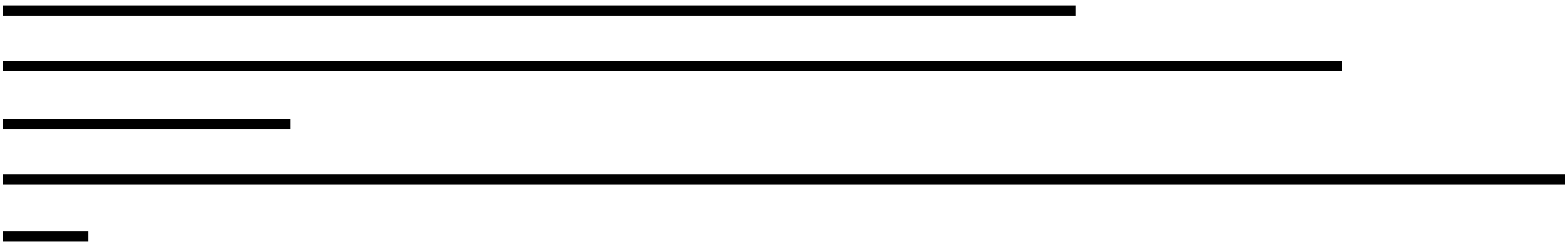
What does an assembly look like?

Metagenome assembly



Scale: 2000 bases

Metagenome assembly

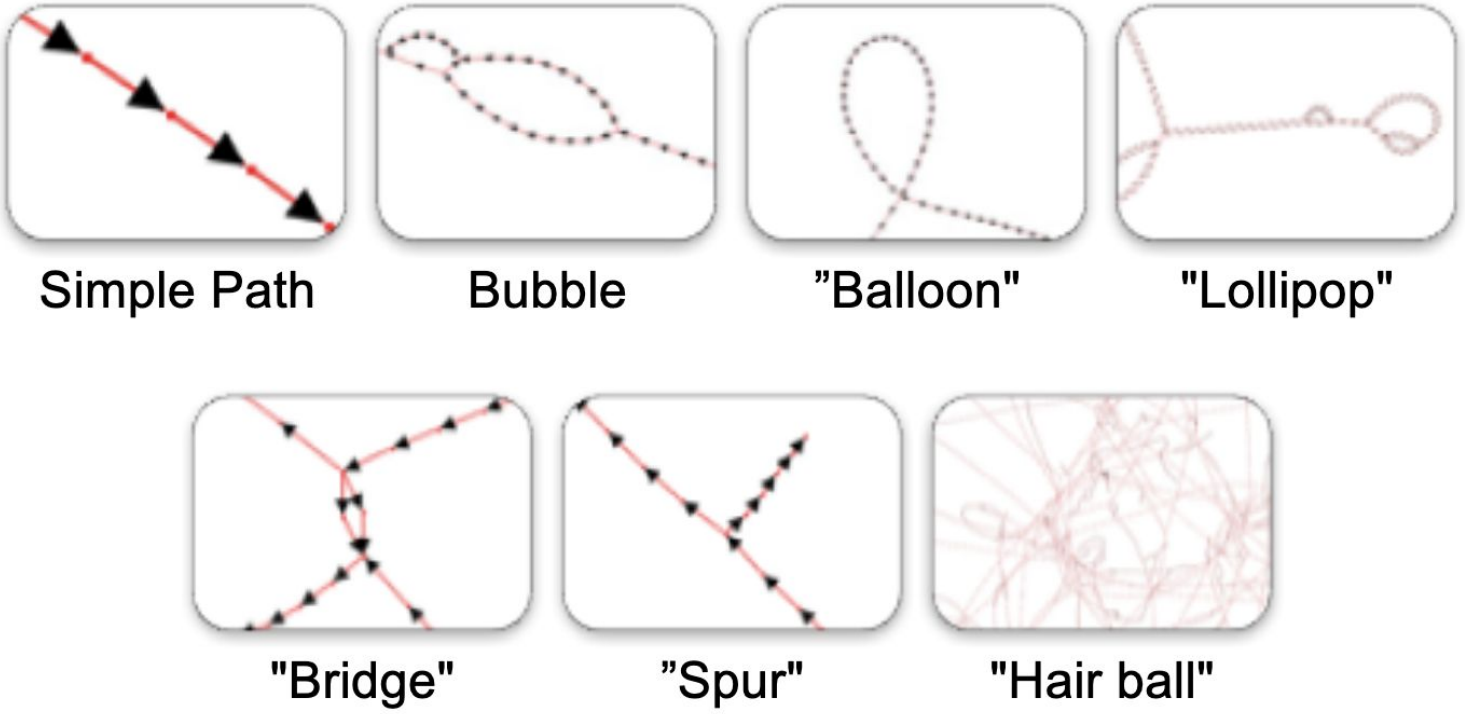


>SRR8859675_0
TATTAATTGGAGCCGTAGCTTCAGAGAAAATACCAGCTACGGTCTTTTCATTTTTATTAGAATTATCGTC
TATACTAACGTCATTTAACTTTTCAGTAATTGCCTTCTCTACTTGTCTAATAAAATTAAATTTATCTTGA
TTTAAACCACTAATTTCCATTAAATTTTCTACTATATTTTTAACCTTTTCTGCGTCATCTGTATTTAAAA
GATTTTTAATATTTGATTTTAGTACAGTTTTCATTCCTCTAAGAGTTGTAATCGTATCTGCTTCATTGTA
AAATTTCTCTGAAAATTCTTCTATTTGCATTTTTCCTCCTAAGAAAACGTTAATTTGCCATTAATTGATA
TACATTCCCCATTATAATTGTATATTTTTTGGTTTGTACTTGCAAGTATGAATTCATTATCAGTCTTTAT
TGATTCTTGATTTAATAATGCATCATATTTGCCACATTTTAAATATTCGTATCCTTTTATATCTGCACAT
>SRR8859675_1
GCGGCAGGCATACCAGCTCGACATATCATAACGAGAGATACTTGTCCTCCAGGTCAGCAGTCAGCAGAAC
AATGGGTTCGGCAACGTTTTGAACAAACGCTGAAGACATTCAGAAGCAAGCATGGGCAGGGTCGGAAGAT
ATGCCTTATCGTGATGATCGATGCTGATCGTCATACCCCTGAAGAACGCAGGAAACAGTTACAGAAGAAT
ATAAAAAGAGAAAACGGAGAGCCGATTGGAATTTTTGTTCCAGCGAGAAACATCCAGAGCTGGATGGCCT

Scale: 2000 bases

When does assembly fail?

When does assembly fail?



https://files.speakerdeck.com/presentations/bca9c71d44b3428090df3d079b48174b/SMRTBifx_09_2016_for_PDF.pdf @infoecho Jason Chin

When does assembly fail?

- Low coverage

How do we evaluate an assembly?

How do we evaluate an assembly?

- Length of contigs

How do we evaluate an assembly?

- Length of contigs

Longer isn't always better

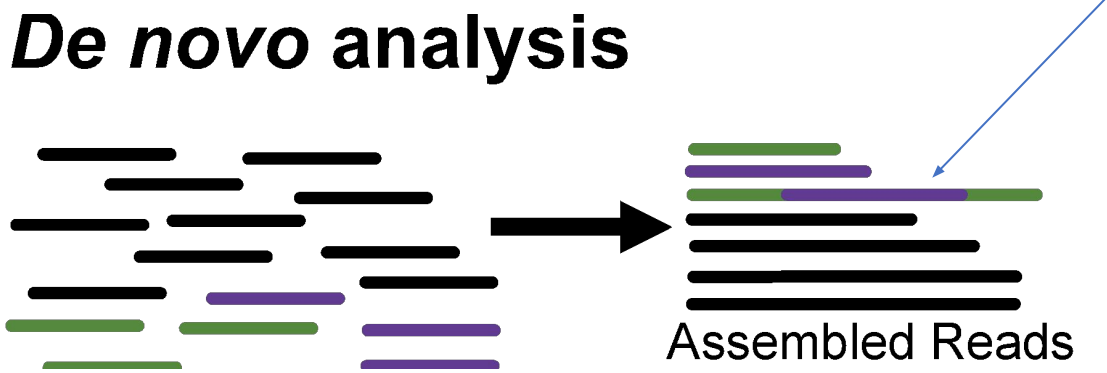


Image: Reiter, T.E., Brown, C.T. *Nat Microbiol* **7,** 193–194 (2022). https://doi.org/10.1038/s41564-021-01027-2
Statement: Mende et al. *Plos ONE* **7**(2): e31386 (2012) https://doi.org/10.1371/journal.pone.0031386

How do we evaluate an assembly?

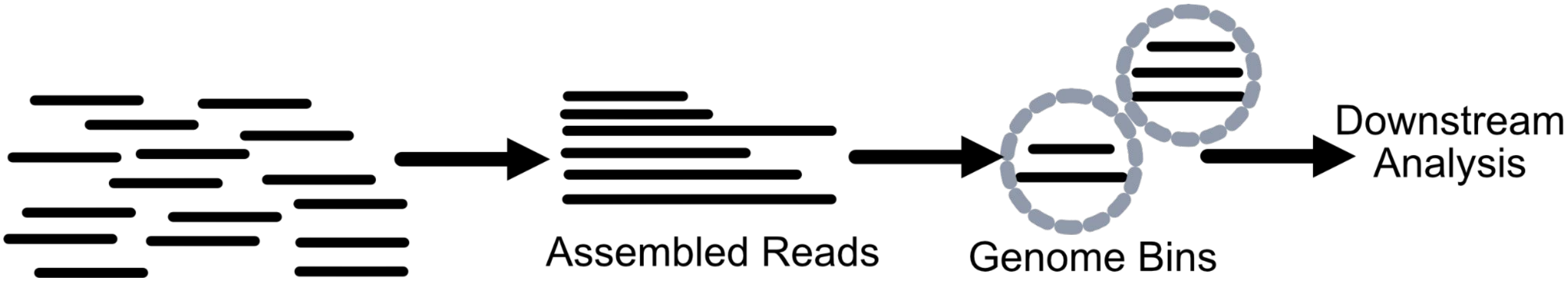
- Length of contigs
- Sequence/read recruitment
  - What fraction of reads map back to the assembly?
  - What fraction of k-mers from the reads are in the assembly

Binning

We have an assembly. Now what? May we haz genomes?

Everything’s made up and the points don’t matter

Assembly & Binning



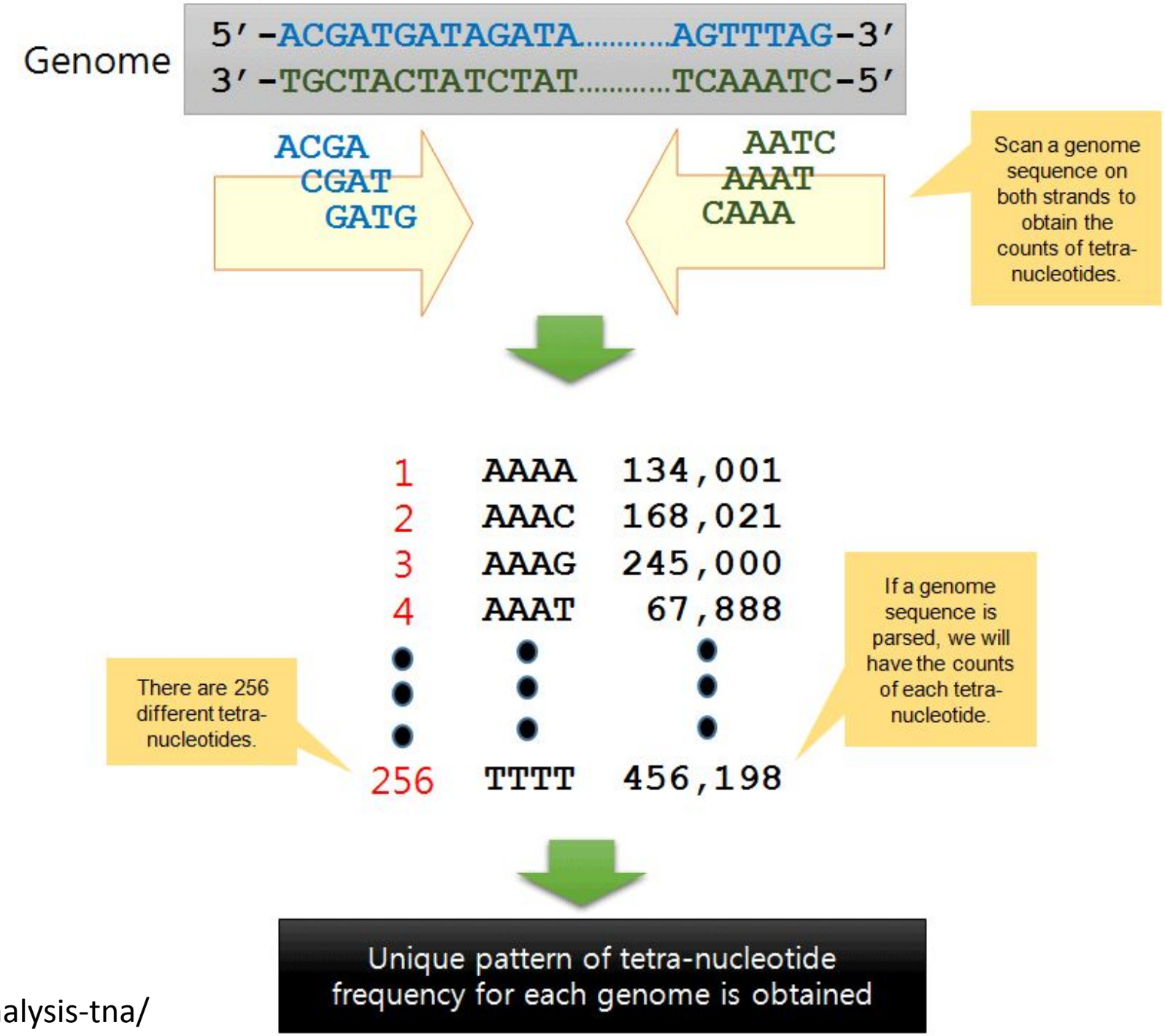
How do we bin?

How do we bin?

- K-mers
- Coverage/abundance
- Single copy marker genes

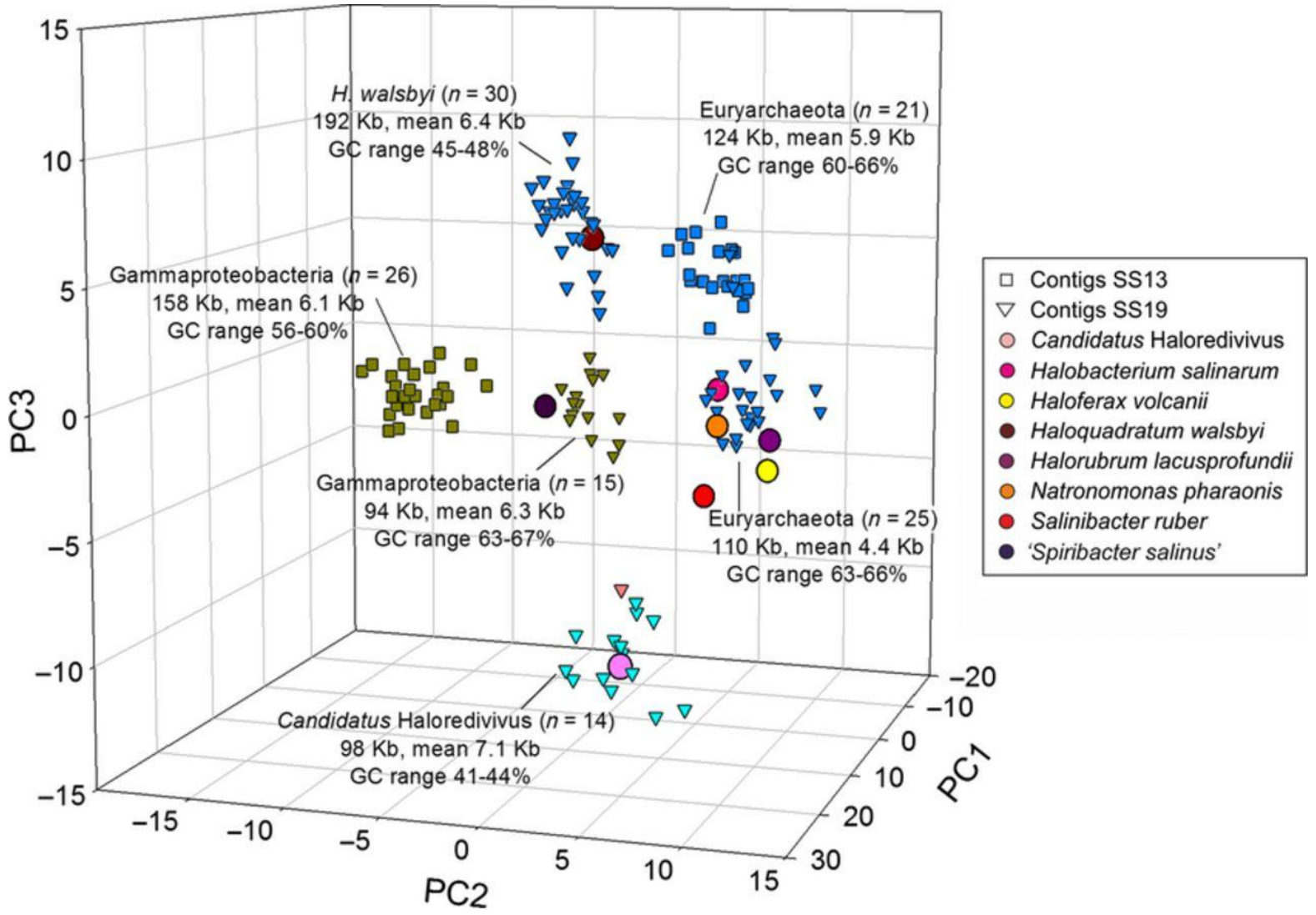
- (pick two, evaluate with the third)

Tetranucleotide frequency



https://help.ezbiocloud.net/tetra-nucleotide-analysis-tna/

Tetranucleotide frequency

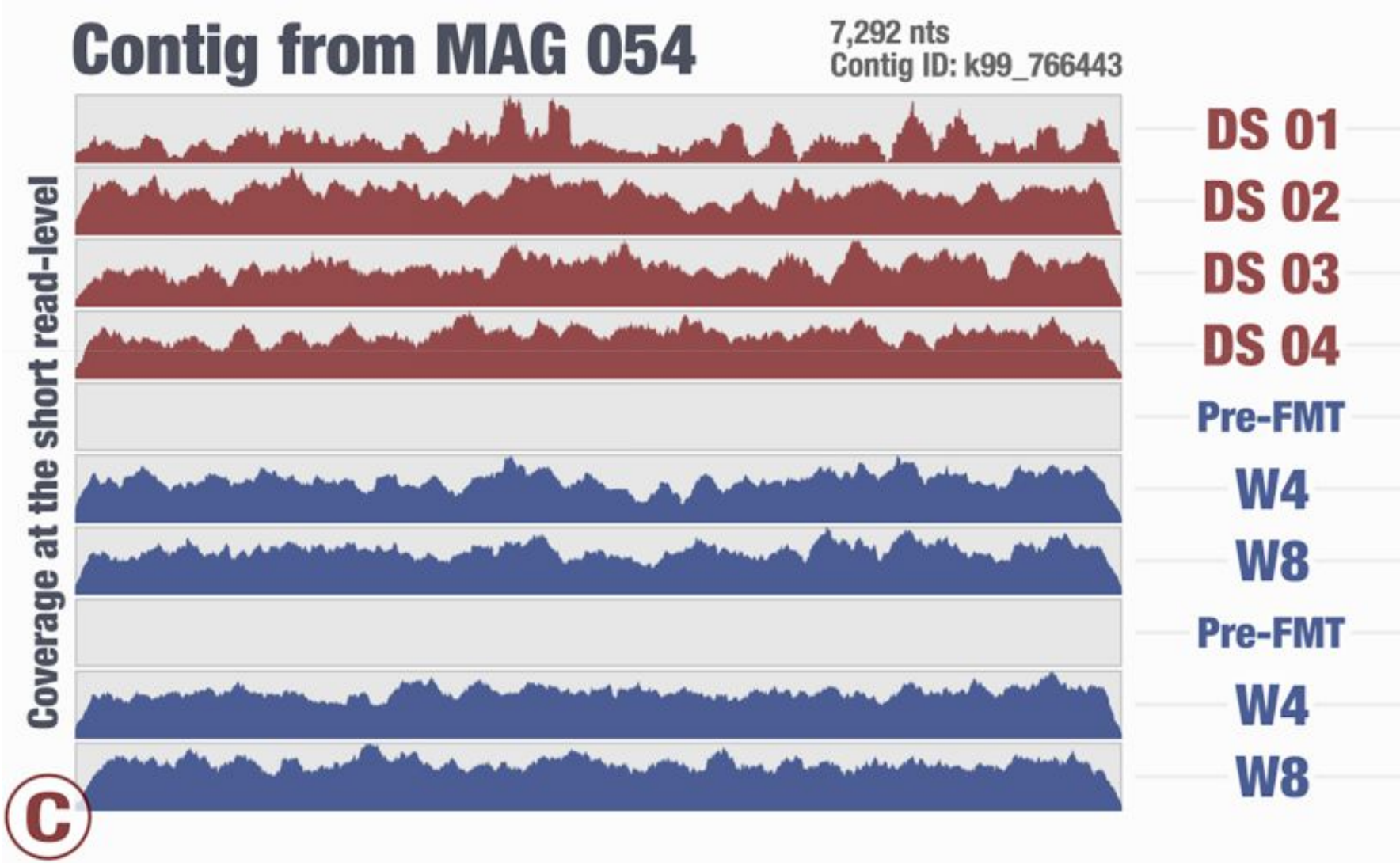


Fernández et al. *FEMS Microbiol Ecol*, **88**:3 (2014) https://doi.org/10.1111/1574-6941.12329

Coverage

https://merenlab.org/2019/11/25/visualizing-coverages/

Coverage in real life



https://merenlab.org/2016/12/14/coverage-variation/

How do we evaluate binning?

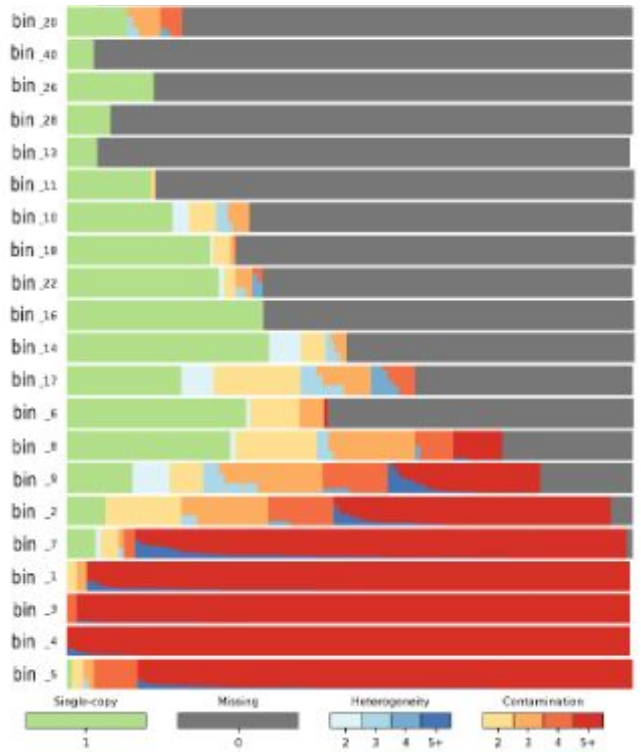
Why do we evaluate binning?

How do we evaluate bins?

- K-mers
- Coverage/abundance
- Single copy marker genes

- (pick two, evaluate with the third)

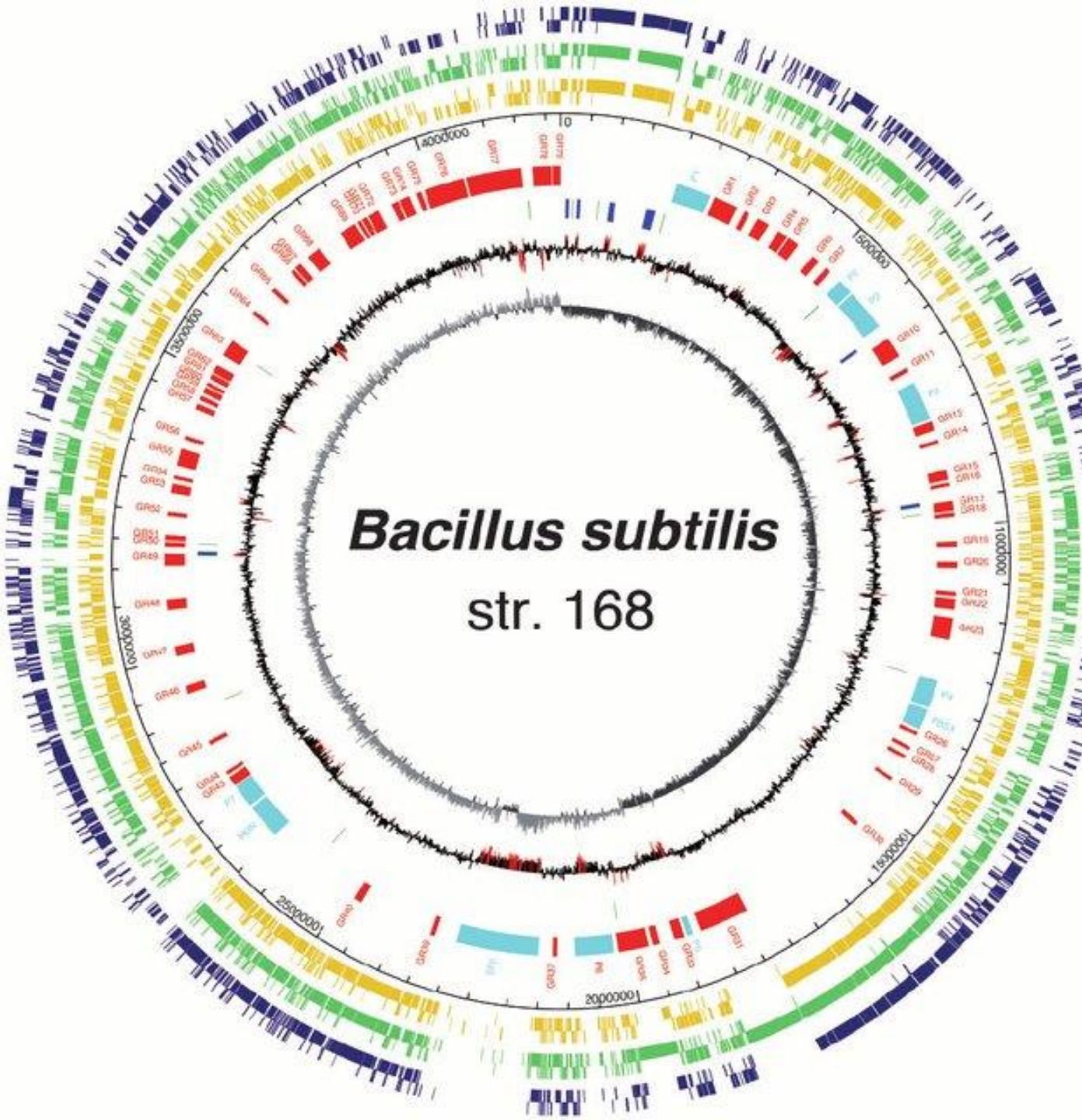
Using single copy marker genes to evaluate bins: checkM



https://kbase.us/applist/applist/apps/kb_Msuite/run_checkM_lineage_wf/release

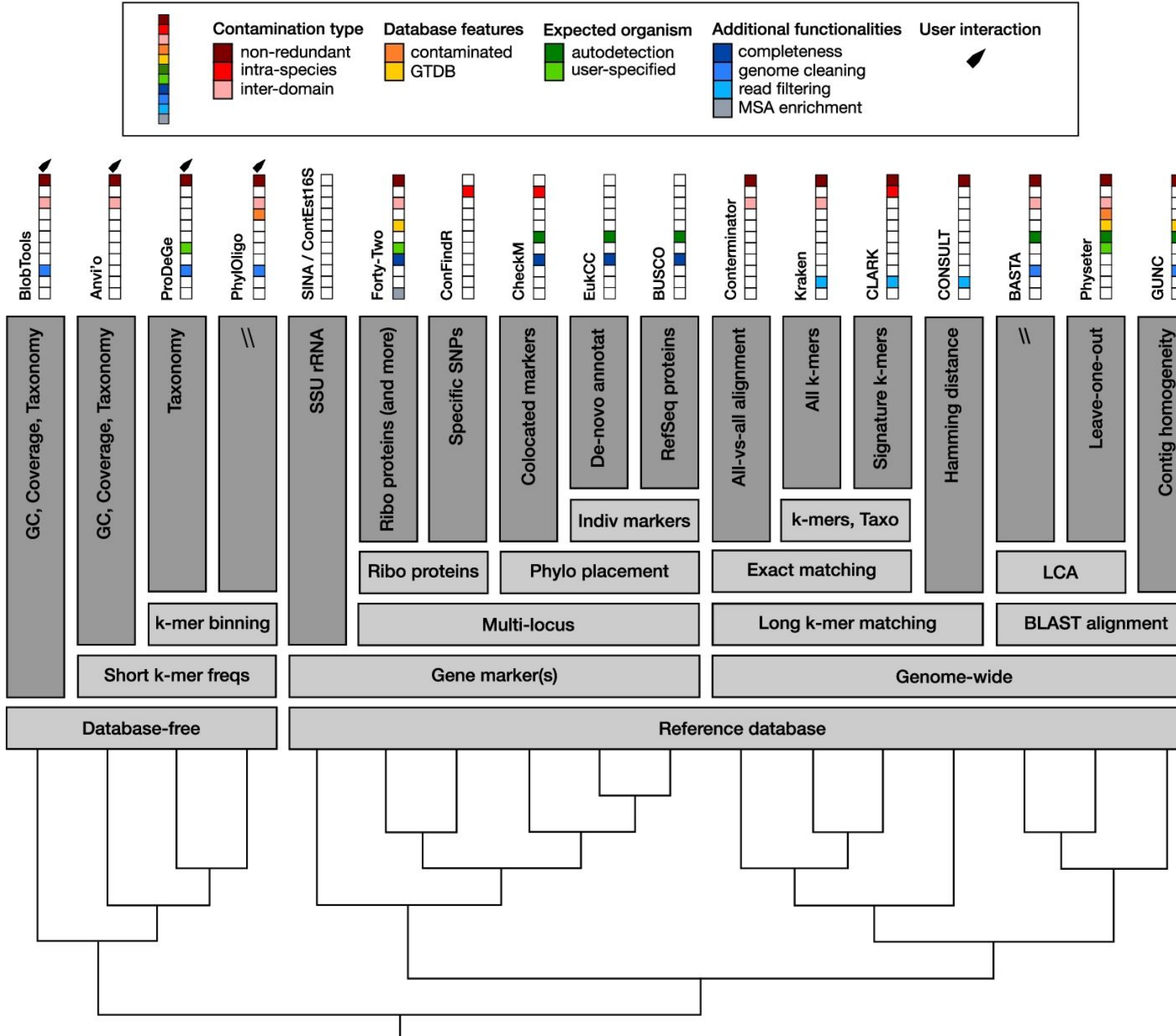
https://www.biostars.org/p/393817/

Using single copy mark genes to evaluate bins: checkM



Barbe et al. *Microbiology*. **155**:6. (2009)

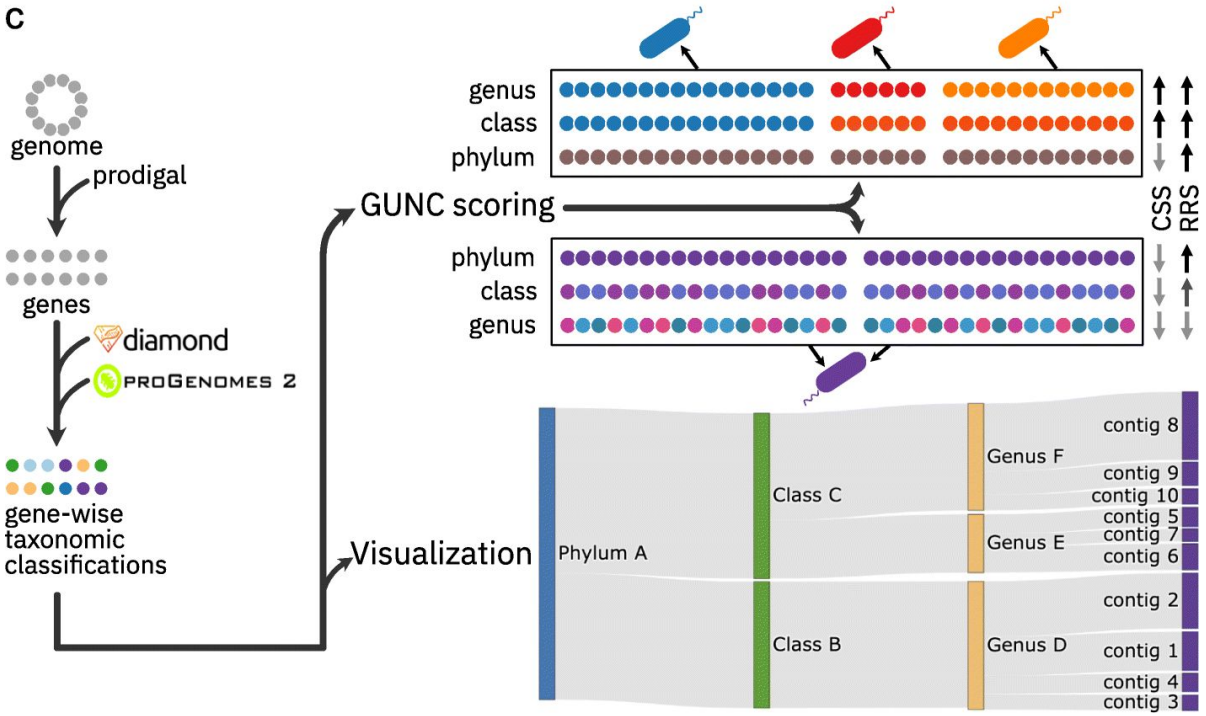
Other ways to evaluate contamination



https://doi.org/10.1186/s13059-022-02619-9

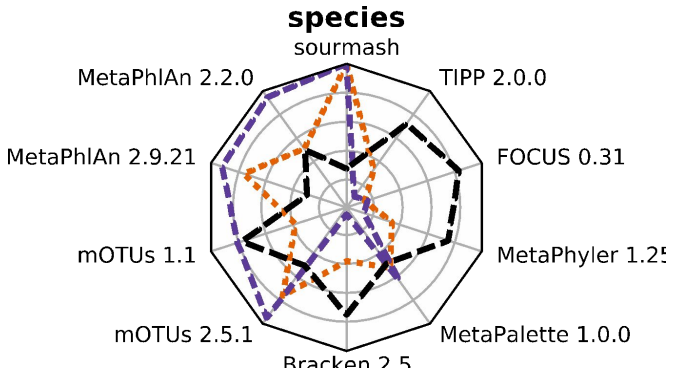
Methods we think are promising to evaluate and *remove* contamination

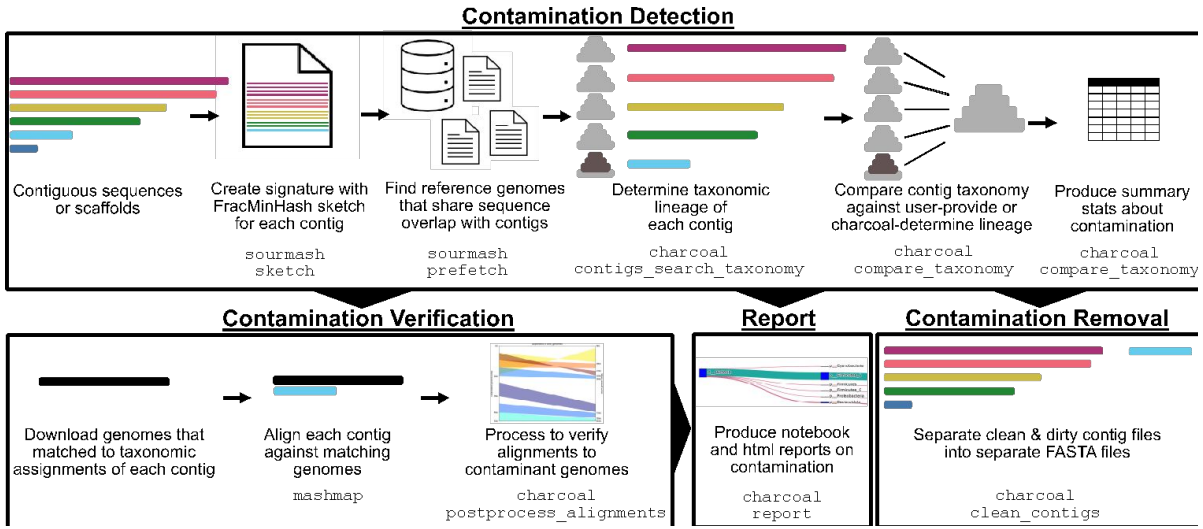
GUNC



Orakov *et al. Genome Biol* **22,** 178 (2021). https://doi.org/10.1186/s13059-021-02393-0

Charcoal





github.com/dib-lab/charcoal

Irber *et al. bioRxiv* (2022). https://doi.org/10.1101/2022.01.11.475838

Where does binning fail most frequently?

Where does binning fail most frequently?

- Plasmids
- Genomic islands (e.g. mobile genetic elements)
- Short contigs

Maguire et al. (2020) *Microbial genomics*, 6(10) https://doi.org/10.1099/mgen.0.000436

 Tutorial 

https://hackmd.io/mo5CbK_XT3CZWuEq8t0Vqw

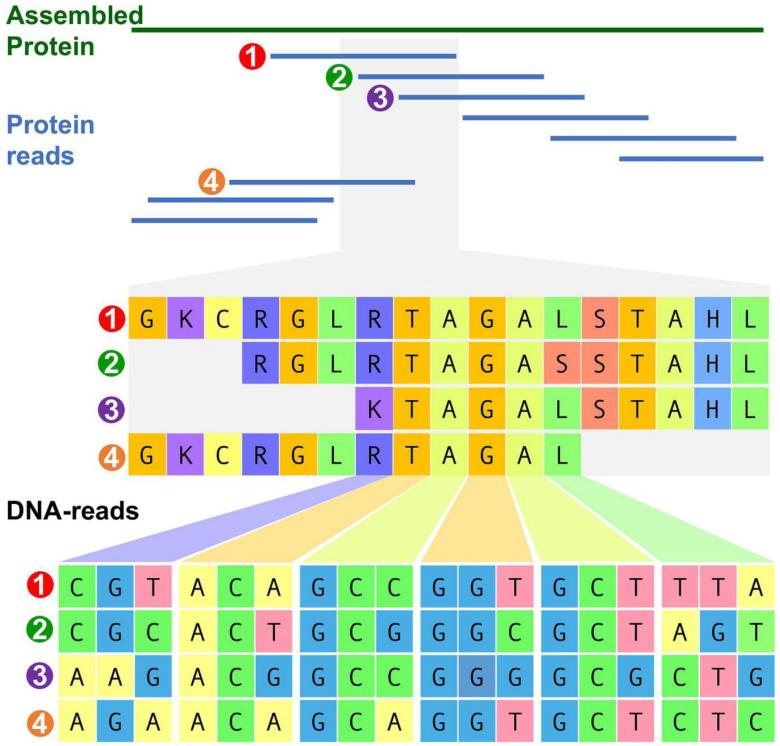
Strategies for when assembly fails

How many of reads did not assemble?

Assemble in protein space

Pros

- With less microdiversity in protein space -> more assembly



Cons

- Combinatorial
- only get proteins not genomes or relationships between genomes

Steinegger et al. *Nat Methods* **16,** 603–606 (2019). https://doi.org/10.1038/s41592-019-0437-4

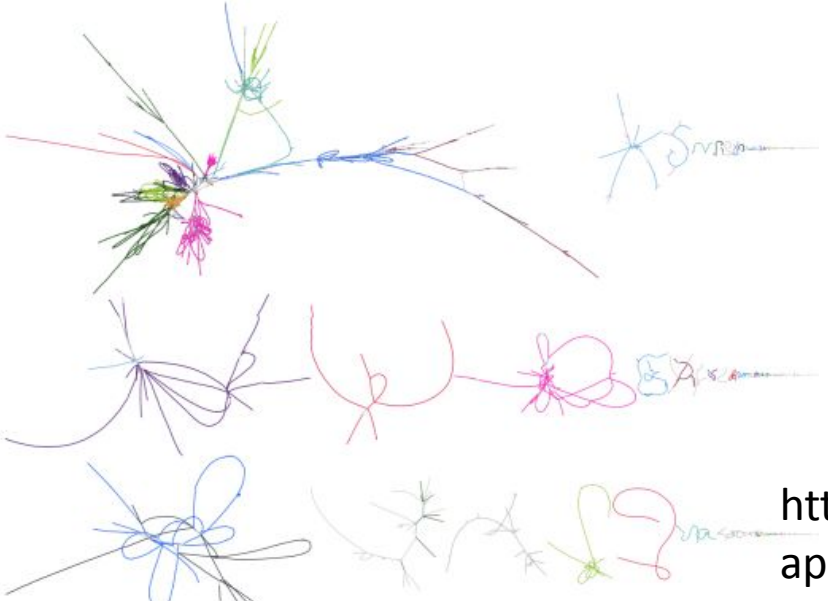
Metagenome assembly graph analysis

Pros

- Contains all the reads
- Mostly organized similarly to how those sequences occur in a genome

Cons

- Messy
- Big
- Tool space is still developing



https://tylerbarnum.com/2018/02/26/how-to-use-assembly-graphs-with-metagenomic-datasets/

Metagenome assembly graph analysis

- **I want to build and analyze assembly graphs myself**
  - https://spacegraphcats.github.io
- **I want to do the same thing you did even though you didn't tell me about it today**
  - https://github.com/dib-lab/2022-dominating-set-differential-abundance-example
- **I want to learn more about biological results that come from using assembly graphs**
  - https://www.biorxiv.org/content/10.1101/2022.06.30.498290v1
  - https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02066-4
- **I want to see more assembly graphs in action**
  - https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02066-4
  - https://www.biorxiv.org/content/10.1101/2022.06.27.497795v1